Chapitre 1

Statistiques descriptives

L'activité statistique comprend deux aspects complémentaires : la *statistique descriptive*, partie indispensable de tout projet d'analyse de données; et *l'inférence statistique*, qui suit souvent — mais pas toujours — l'analyse descriptive. Un sondage, une enquête, une recherche scientifique produisent de nombreuses données qu'on ne peut déchiffrer tant qu'on ne les a pas réduites à des dimensions compréhensibles. Supposons, par exemple, que l'administrateur d'une petite municipalité rassemble des données recueillies auprès des 5000 ménages de la ville — des données sur le nombre de personnes qui y vivent, le nombre d'enfants, la consommation d'eau, etc. Le résultat est un tableau de 5000 lignes et autant de colonnes que de questions posées. Les techniques qui permettent de réduire la dimension de ces données — les tableaux, les graphiques, ainsi que les « mesures statistiques » telles les moyennes et les ratios — font partie de la statistique descriptive.

Si la ville ne comprend que ces 5000 ménages, et si seuls ceux-là intéressent l'administrateur, alors l'étude est un *recensement*, et les 5000 ménages constituent ce qu'on appelle une *population* : c'est la totalité des objets d'intérêt. Dans ce cas les analyses descriptives suffisent. Mais le coût d'un recensement complet de la population est rarement abordable. Il est nécessaire alors de limiter l'étude à un *échantillon*, une partie de la population. Les données qui en résultent seront quand même soumises à une analyse descriptive : qu'il s'agisse d'un échantillon ou d'une population, il faut que les données soient résumées. Mais l'étude ne peut s'arrêter là, car son objet, c'est la population, pas l'échantillon. L'échantillon n'est qu'une image de la population, une image qu'on souhaite fidèle, mais qui n'est jamais parfaite. Il révèle — à peu près — certaines des caractéristiques de la population, mais il peut aussi « révéler » des choses qui ne sont pas vraies. Si l'échantillon montre qu'il y a plus d'enfants par ménage dans le secteur Sud que dans le secteur Nord, on est en droit de se demander si ce constat est vrai aussi de la population ou s'il s'agit d'un accident du hasard. Dans quelle mesure peut-on faire confiance aux observations faites sur un échantillon et déduire qu'elles sont également vraies (ou à peu près) de la population ? C'est à ces questions que doivent répondre les techniques d'inférence statistique.

Des questions que nous n'aborderons formellement qu'au chapitre 6, lequel sera précédé de trois chapitres consacrés à la *théorie des probabilités*, base théorique de l'inférence statistique. Dans ce chapitre et le suivant, cependant, nous nous concentrons sur la statistique descriptive et donc nous n'y ferons pas la distinction — très importante, par ailleurs — entre *population* et *échantillon*.

1.1 Introduction : variables et distributions

Pour commencer, un peu de vocabulaire. Le tableau A.1 en annexe (dont une partie est reproduite ci-dessous) présente une série de données concernant des professeurs d'université. L'ensemble des professeurs constitue la *population*. Les membres d'une population sont appelés des *unités statistiques*, ou simplement des *unités*. Chaque ligne du tableau représente une unité, identifiée par un numéro (qui peut aussi être un nom) dans la première colonne. Chaque colonne représente une *variable*, et les nombres ou lettres qui y figurent sont les *valeurs*, ou *modalités* de la variable. La variable « Sexe », par exemple, a pour modalités les lettres *F* et *M* ; les valeurs de la variable « Date d'entrée » sont des entiers de 1980 à 2012. Les valeurs de la variable « Département » sont les noms des différents départements de l'université.

On distingue deux catégories de variable: les variables *quantitatives*, comme les salaires et le nombre de mois d'expérience, sont celles dont les modalités sont des quantités; et les variables *qualitatives*, dont les modalités sont des caractéristiques non mesurables, comme le département et sexe.

Extrait du tableau A.1 - *Quelques données sur un groupe de professeurs*

Identité	Sexe	Date d'entrée	Département	Salaire à l'entrée	Salaire en 2012	Expérience
1	F	1995	Management	16 598	109 268	22
2	M	1984	Management	9 386	134 244	27
3	F	2008	Management	34 446	81 170	22
4	M	1990	Sc. économiques	15 962	159 532	30
5	M	1999	Marketing	23 413	153 600	20
6	M	1995	Sc. comptables	19 838	140 175	28
7	M	2007	Management	34 541	107 395	21
8	F	2001	Finance	30 797	126 751	48
9	M	2004	Sc. comptables	20 726	109 893	22
10	M	1990	Finance	13 038	126 751	28
11	M	2007	Sc. comptables	30 005	91 786	21
…	…	…	…	…	…	…

On peut résumer les caractéristiques d'une variable en énumérant ses modalités et en indiquant la fréquence de chacune dans la population. Le tableau 1.1.1 résume les caractéristiques de la variable « Département ». L'effectif correspondant à une valeur donnée *x*, c'est le nombre d'unités pour lesquelles la variable prend la valeur *x*. *L'effectif total n* est le nombre d'unités dans la population (ou l'échantillon). La *fréquence* d'une valeur est l'effectif de cette valeur divisé par *n*. La somme des fréquences est nécessairement égale à 1. La fréquence est parfois multipliée par 100, de façon à représenter un pourcentage. Cette correspondance entre les valeurs d'une variable et les effectifs ou fréquences correspondantes est appelée *distribution*.

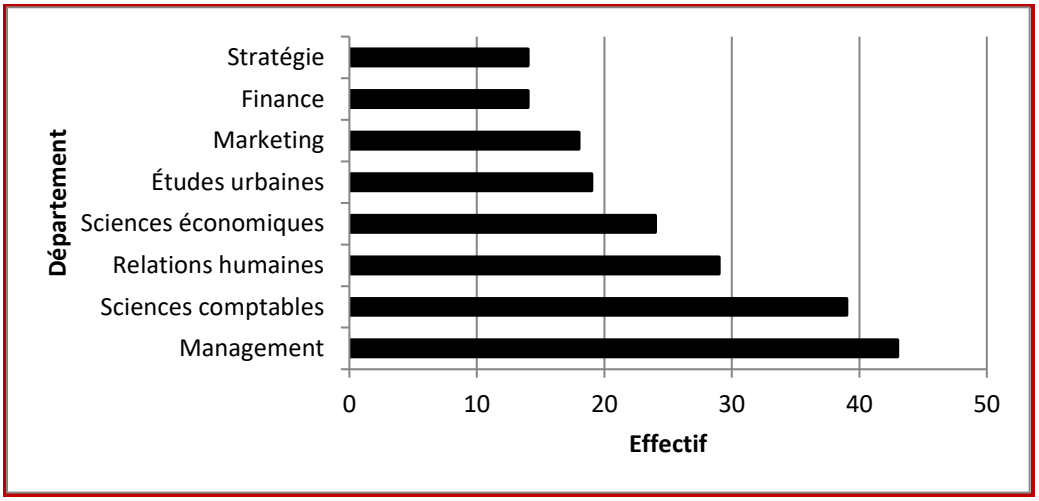
Tableau 1.1.1

Distribution de la variable « Département »
Données du tableau A.1

Valeurs	Effectif	Fréquence
Études urbaines	19	0,095
Finance	14	0,070
Management	43	0,215
Marketing	18	0,090
Relations humaines	29	0,145
Sciences comptables	39	0,195
Sciences économiques	24	0,120
Stratégie	14	0,070
	200	1

La figure 1.1.1 présente un graphique de cette distribution, appelé diagramme à barres. Les barres sont placées en ordre croissant de haut en bas, mais d'autres arrangements sont possibles.

Figure 1.1.1
Présentation graphique de la distribution de la variable « Département »
Données du tableau 1.1.1



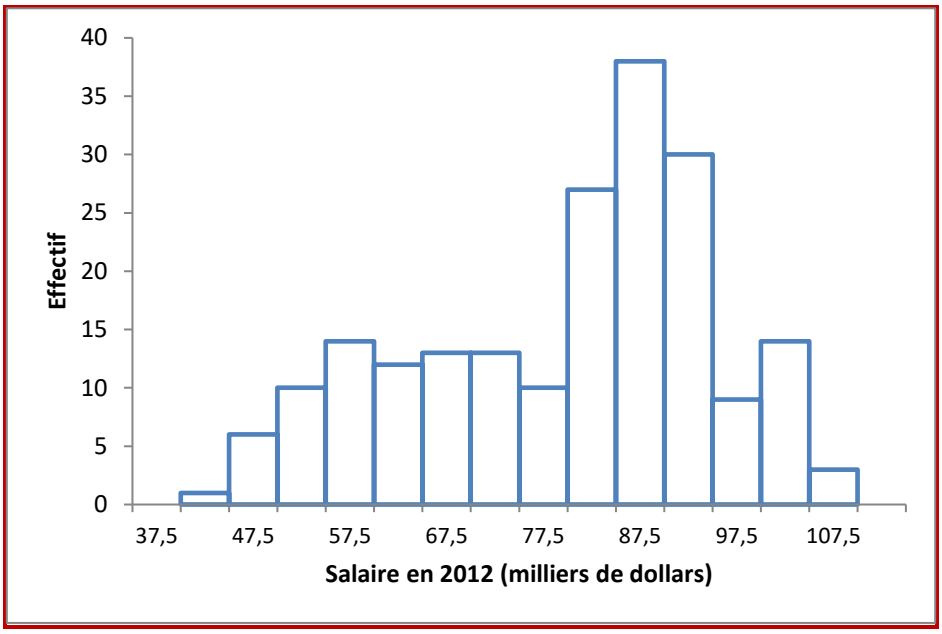
Parfois les valeurs d'une variable sont si nombreuses (et les effectifs en conséquence très faibles et souvent nuls) que le groupement des valeurs est inévitable. C'est le cas de la variable « Salaire en 2012 » de la population présentée au tableau A.1. Le tableau 1.1.2 présente une distribution dans laquelle les classes sont les intervalles]35 000 ; 40 000] ,]40 000 ; 50 000], ... ,]110 000 ; 115 000]. Les centres de ces intervalles, appelés *points-milieux*, figurent dans la deuxième colonne du tableau. L'effectif de la 2e classe,]50 000 ; 55 000], est 10, le nombre de professeurs dont le salaire est supérieur à 50 000 $ et inférieur ou égal à 55 000 $.

Tableau 1.1.2
Distribution de la variable « Salaire en 2012 »
Données du tableau A.1

Intervalle	*Point-milieu*	*Effectif*	*Fréquence*
]35000 ; 40000]	37500	0	0,000
]40000 ; 45000]	42500	1	0,005
]45000 ; 50000]	47500	6	0,030
]50000 ; 55000]	52500	10	0,050
]55000 ; 60000]	57500	14	0,070
]60000 ; 65000]	62500	12	0,060
]65000 ; 70000]	67500	13	0,065
]70000 ; 75000]	72500	13	0,065
]75000 ; 80000]	77500	10	0,050
]80000 ; 85000]	82500	27	0,135
]85000 ; 90000]	87500	38	0,190
]90000 ; 95000]	92500	30	0,150
]95000 ; 100000]	97500	9	0,045
]100000 ; 105000]	102500	14	0,070
]105000 ; 110000]	107500	3	0,015
]110000 ; 115000]	112500	0	0,000
		200	1

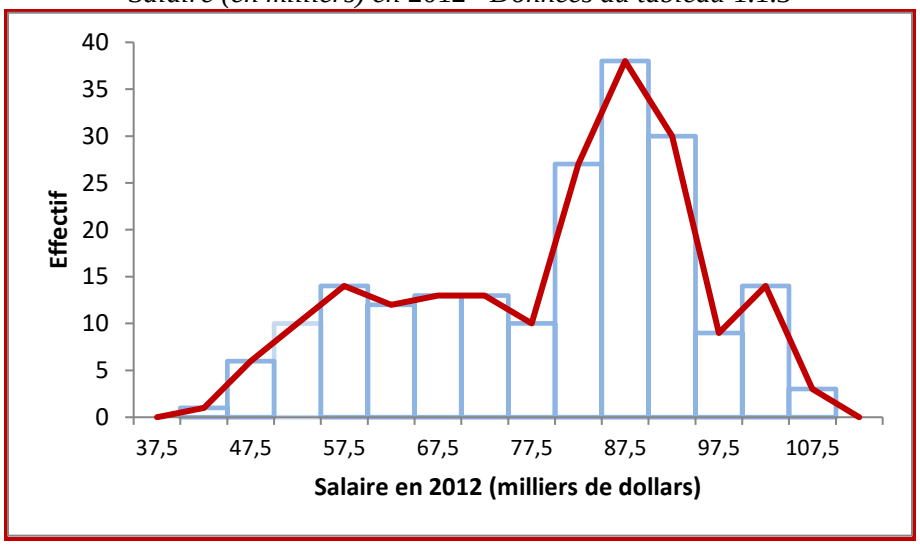
L'histogramme est la représentation la plus courante de la distribution d'une variable quantitative (voir la figure 1.1.2).

Figure 1.1.2
Histogramme de la distribution du salaire de 2012
Données du tableau 1.1.2



Mais une distribution peut également être présentée à l'aide d'un polygone des fréquences, comme dans la figure 1.1.3 (l'histogramme en arrière-plan ne fait pas normalement partie du graphique.)

Figure 1.1.3
Polygone des fréquences
Salaire (en milliers) en 2012 - *Données du tableau* 1.1.3



Un polygone des fréquences a l'avantage de faciliter les comparaisons de distributions, comme dans la figure 1.1.4, qui présente la distribution des salaires pour deux sous-populations: les femmes et les hommes (tableau 1.1.3). Note z bien qu'une comparaison de distributions ne se fait qu'avec les fréquences et non avec les effectifs, à moins que les effectifs des deux populations soient les mêmes.

Tableau 1.1.3
Distribution de la variable « Salaire de 2012»
Données tirées du tableau A.1

Point-milieu	*Femmes*	*Hommes*
]35000 ; 40000]	0,0000	0,0000
]40000 ; 45000]	0,0123	0,0000
]45000 ; 50000]	0,0370	0,0252
]50000 ; 55000]	0,0617	0,0420
]55000 ; 60000]	0,0988	0,0504
]60000 ; 65000]	0,1111	0,0252
]65000 ; 70000]	0,0988	0,0420
]70000 ; 75000]	0,0988	0,0420
]75000 ; 80000]	0,0741	0,0336
]80000 ; 85000]	0,1111	0,1513
]85000 ; 90000]	0,1975	0,1849
]90000 ; 95000]	0,0494	0,2185
]95000 ; 100000]	0,0000	0,0756
]100000 ; 105000]	0,0370	0,0924
]105000 ; 110000]	0,0123	0,0168
]110000 ; 115000]	0,0000	0,0000
	1	1

Figure 1.1.4
Comparaison des salaires - hommes et femmes
*Données du tableau*1.1.3

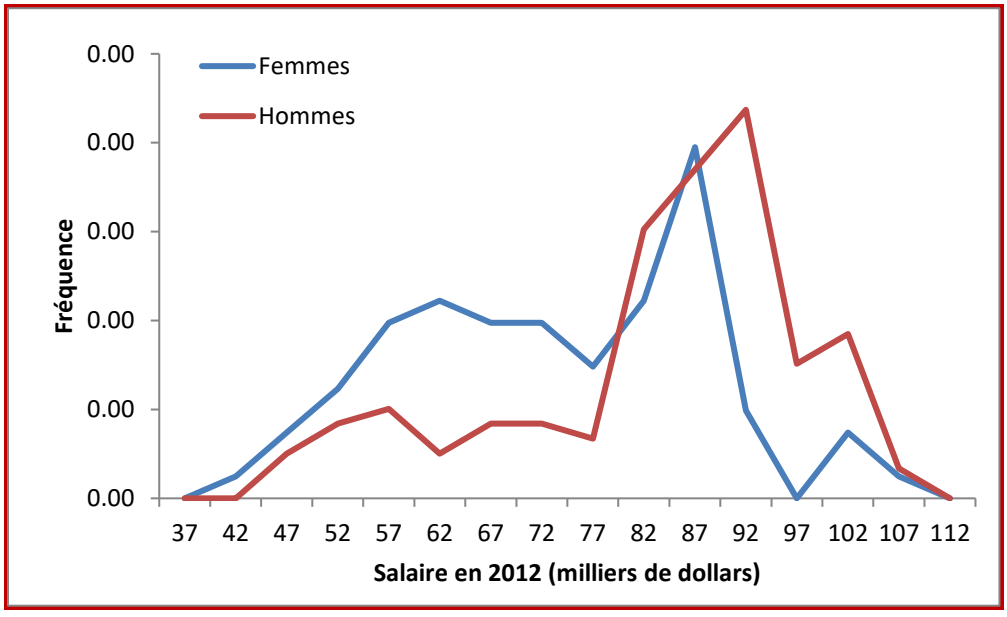
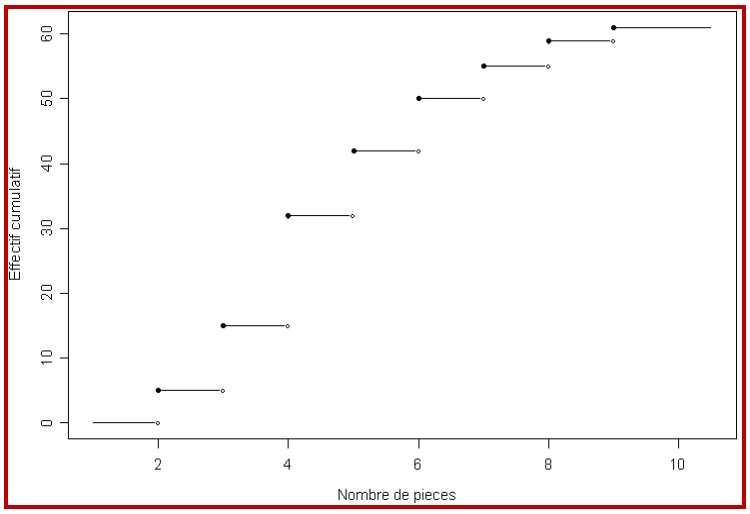


Tableau 1.1.4
Distribution cumulative du nombre de pièces dans un échantillon de 61 *maisons*

Nombre de pièces	*Effectif*	*Effectif cumulé*	*Fréquence cumulée*
1	0	0	0,00
2	5	5	0,08
3	10	15	0,25
4	17	32	0,52
5	10	42	0,69
6	8	50	0,82
7	5	55	0,90
8	4	59	0,97
9	2	61	1,00
10	0	61	1,00

Distribution cumulative
Il est parfois utile d'accumuler les effectifs successifs, de façon à obtenir ce qu'on appelle la ***distribution cumulative***. Le tableau 1.1.4 présente la distribution cumulative du nombre de pièces dans un échantillon de 61 maisons.

Figure 1.1.5
Distribution cumulative du nombre de pièces dans un échantillon de 61 *maisons*



1.2 Mesures de tendance centrale

Une mesure de tendance centrale est un indice de la position d'une série de données ou d'une distribution. Elle donne une idée de l'ordre de grandeur des données. La plus connue des mesures de tendance centrale est la *moyenne arithmétique*, ou *moyenne* tout court. Nous définissons formellement cette mesure, ainsi que deux autres, la *médiane* et le *mode*.

La moyenne arithmétique

Définition *Moyenne d'une série de données*

Soit $x_1, x_2,..., x_n$ une série de n données. Leur *moyenne arithmétique* est définie par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.2.1)$$

La définition suivante s'applique lorsque les données sont présentées sous la forme d'une distribution,

Définition *Moyenne d'une distribution*

Soit une variable X dont les valeurs sont $x_1, x_2, \ldots, x_p$, avec effectifs correspondants $n_1, n_2, \ldots, n_p$ et fréquences $f_1, f_2, \ldots, f_p$. La moyenne arithmétique est définie par

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{p} x_i n_i = \sum_{i=1}^{p} x_i f_i \text{ , où } n = \sum_{i=1}^{n} n_i \qquad (1.2.2)$$

Lorsque les données sont groupées, l'information n'est pas assez détaillée pour permettre le calcul exact de la moyenne. La formule ci-dessus peut néanmoins être appliquée—en remplaçant les valeurs x_i par les points-milieux— et donne une valeur approximative de la moyenne. Le résultat est exact si toutes les valeurs d'une classe, — encore leur moyenne—sont égales au point-milieu de la classe.

Exemple **1.2.1** Considérons les données suivantes :

1 ; 1 ; 1 ; 1 ; 1 ; 1 ; 1 ; 2 ; 2 ; 2 ; 2 ; 3 ; 3 ; 3 ; 4 ; 4

On a $n = 16$, et par la première formule, la moyenne est

$$\bar{x} = \frac{1}{16}\sum_{i=1}^{16} x_i = \frac{1+1+1+1+1+1+1+2+2+2+2+3+3+3+4+4}{16} = 2.$$

Les mêmes données peuvent être présentées sous la forme d'une distribution :

x_i	1	2	3	4	
Effectif n_i	7	4	3	2	16
Fréquence f_i	7/16	4/16	3/16	2/16	1

Une application de la deuxième formule, $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} x_i n_i$ avec $p = 4$ donne

$$\bar{x} = \frac{1\times 7 + 2\times 4 + 3\times 3 + 4\times 2}{16},$$

ce qui équivaut à une réécriture du numérateur qui consiste à remplacer la somme 1+1+1+1+1+1+1 par 1×7, la somme 2+2+2+2 par 2×4, ainsi de suite. Si on effectue la division par 16 terme par terme, on a ceci : $\bar{x} = 1\frac{7}{6} + 2\frac{4}{16} + 3\frac{3}{16} + 4\frac{2}{16}$, une application de la formule $\bar{x} = \sum_{i=1}^{p} x_i f_i$. ■

La médiane

Une autre mesure de tendance centrale est la *médiane*.

Définition *Médiane d'une série de données*

La *médiane* est la donnée centrale d'une série, lorsque les données sont rangées en ordre croissant ou décroissant.

Lorsque les données sont en nombre pair, la médiane est la moyenne des deux données centrales.

Par exemple, la médiane des données

1, 3, 5, **7**, 9, 10, 13

est 7.

La médiane des données

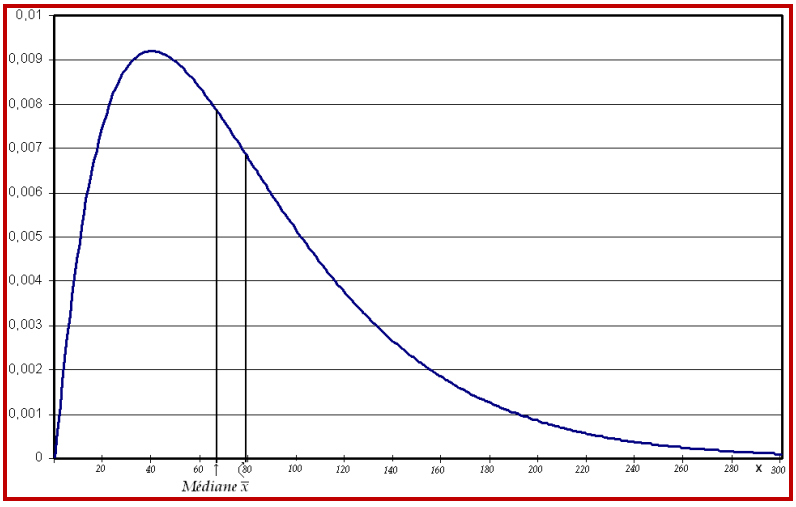
1, 3, 5, **7 , 9**, 10, 13, 15

est (7 + 9)/2 = 8.

Médiane ou Moyenne arithmétique? S'il est vrai que la moyenne arithmétique est la mesure la plus souvent utilisée, la médiane a aussi son utilité et peut parfois pallier certains défauts de la moyenne arithmétique, dont l'un est sa sensibilité aux données extrêmes — le fait que quelques données excentriques peuvent donner à la moyenne une valeur dont on ne peut pas vraiment dire qu'elle représente l'ensemble. Une poignée de gens avec des revenus énormes peut tirer le revenu moyen d'une population vers le haut et donner une impression exagérée de richesse. C'est pour cela, d'ailleurs, que lorsque les instituts de statistiques comme le Bureau de la statistique du Québec et Statistique Canada présentent des données sur les revenus, ils donnent la médiane ainsi que la moyenne. On trouve généralement que la médiane est inférieure à la moyenne arithmétique. La figure 1.2.1 illustre les positions relatives de la moyenne et la médiane dans une population asymétrique avec concentration à gauche : certaines valeurs sont bien au-dessus de la majorité de la population qui, elle, forme un sommet à gauche. Il est clair que lorsque la concentration des données est à droite, c'est la moyenne arithmétique qui est inférieure à la médiane. Lorsque la population est de forme *symétrique*, les deux mesures coïncident.

Figure 1.2.1

Relation entre la médiane et la moyenne arithmétique dans une population asymétrique avec concentration à gauche

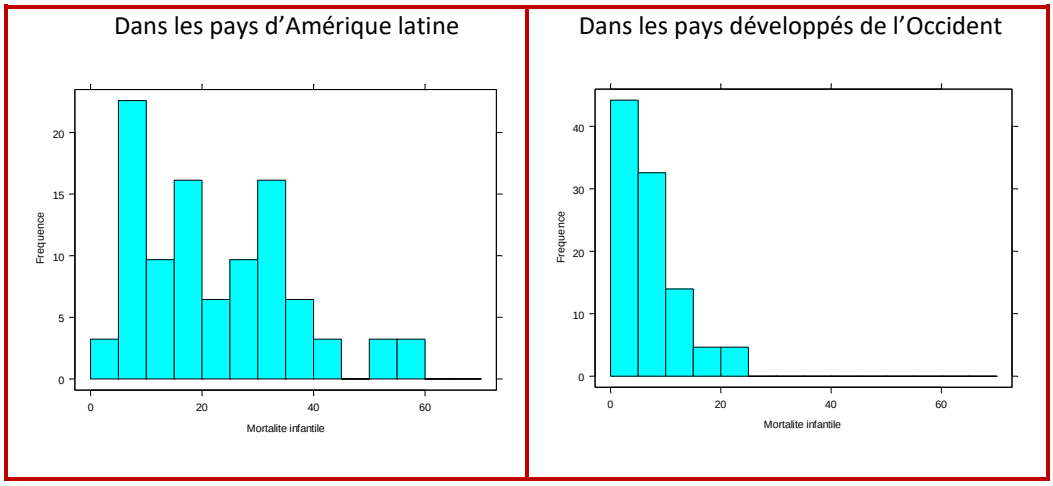


1.3 Mesures de dispersion

La description sommaire que fournissent la moyenne et la médiane cache beaucoup de choses. Entre autres, elle ne fournit aucune information sur la *dispersion* des données, une caractéristique particulièrement importante dans certains contextes. Par exemple, deux classes à l'université pourraient avoir la même moyenne à un examen tout en étant fondamentalement différentes. L'une pourrait être constituée d'un mélange d'étudiants très forts et d'étudiants très faibles alors que l'autre pourrait rassembler surtout des étudiants moyens. Remarquez aussi les deux histogrammes de la figure 1.3.1. Ils représentent les taux de mortalité infantile[1] dans des pays d'Amérique latine et dans des pays occidentaux développés. On voit bien que le taux est plus élevé *en moyenne* dans les pays d'Amérique latine (7,88 dans les pays occidentaux ; 22,74 en Amérique latine). Mais on voit aussi que les taux sont plus *dispersés* en Amérique latine : certains pays ont des taux très élevés, d'autres plutôt bas. Les écarts entre les pays occidentaux sont beaucoup plus petits. C'est cette dispersion, aisément perçue dans un graphique, que nous cherchons à *mesurer*—tout comme on mesure la tendance centrale—de façon à transmettre l'information à l'aide d'un seul chiffre.

Figure 1.3.1

Distributions des taux de mortalité infantile



Écart-type et variance

Comme mesure de la dispersion d'une série de données, on utilise l'*écart-type*; une quantité associée à l'écart-type est la *variance*.

Définition *Variance et écart-type d'une série de données*

Soit $x_1, x_2, \ldots, x_n$ une série de données. La variance σ^2 de la série est donnée par

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}. \qquad (1.3.1)$$

L'*écart-type* est la racine carrée de la variance, soit

[1] Nombre de décès par 1000 naissances.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}. \qquad (1.3.2)$$

Lorsqu'on effectue ces calculs pour les pays d'Amérique latine et pour les pays de l'Occident (figure 1.3.1), on a les résultats suivants :

Amérique latine : $\sigma = 13{,}6$; pays occidentaux : 4,79

Remarque *La notion d'écart-type tire son importance de certaines propriétés qui seront développées plus tard. Pour l'instant nous la traitons comme simple mesure descriptive, donc une quantité qui, pour être utile, doit avoir un sens immédiat. Ce qui n'est pas toujours évident : que signifie la valeur* $\sigma = 13{,}6$ *de l'écart-type des taux de mortalité des pays d'Amérique latine ? Toute seule, pas grand-chose. Elle prend son sens lorsqu'on la compare à l'écart-type* $\sigma = 4{,}79$ *pour les pays de l'Occident : ces derniers diffèrent moins entre eux que les premiers. En fin de compte, la valeur numérique d'un écart-type n'a de sens que par comparaison à un ou plusieurs autres. Par exemple, un professeur qui calcule systématiquement l'écart-type des résultats d'un examen finit par se faire une idée de ce qu'est un écart-type normal : environ* 20 %*. Il ou elle saura donc qu'il s'est produit une anomalie le jour où l'écart-type est de* 10 *ou de* 30*. Notons en passant qu'une simple moyenne n'est guère plus facile à interpréter dans un contexte non familier. Comment réagir au fait que les Québécois ont acheté* 30,0 *chopines de fraises cette année ? À moins que vous ayez un intérêt particulier pour la consommation de fraises au Québec, cette information vous laissera froid.* ■

Variance d'une distribution

Lorsque les données sont présentées sous forme de distribution, la variance (ou l'écart-type) peut quand même être déterminée à l'aide des formules suivantes dérivées en développant les données comme à l'exemple 1.2.1.

Définition *Variance et écart-type d'une distribution*

Soit X une variable dont les valeurs sont $x_1, x_2, \ldots, x_p$, avec effectifs correspondants $n_1, n_2, \ldots, n_p$ et fréquences $f_1, f_2, \ldots, f_p$. La variance et l'écart-type sont alors définis respectivement par

$$\sigma^2 = \sum_{i=1}^{p}(x_i - \bar{x})^2 f_i \quad \text{et} \quad \sigma = \sqrt{\sum_{i=1}^{p}(x_i - \bar{x})^2 f_i} \qquad (1.3.3)$$

Lorsque les données sont groupées, on peut employer les formules ci-dessus en remplaçant les valeurs x_i par les points-milieux. On obtient ainsi une *approximation* de la variance et de l'écart-type.

Formules de calcul :

On démontre facilement que

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}{n} = \frac{\sum_{i=1}^{n}x_i^2}{n} - \bar{x}^2 \qquad (1.3.4)$$

et que

$$\sum_{i=1}^{p}(x_i - \bar{x})^2 f_i = \sum_{i=1}^{p}x_i^2 f_i - \bar{x}^2 \ . \qquad (1.3.5)$$

Ces formules sont traditionnellement présentées comme une façon de faciliter le calcul manuel de la variance, un avantage peu important étant donné les facilités de calculs actuels. Elles restent néanmoins utiles dans certaines démonstrations mathématiques.

Coefficient de variation

Nous avons déjà signalé que pour interpréter un écart-type il faut le comparer à quelque chose : l'écart-type d'une autre population, par exemple, ou l'écart-type de la même population à un autre moment. Mais des fois les écarts-types bruts ne sont tout simplement pas comparables comme tels : ils doivent d'abord être normalisés d'une certaine façon. Si on compare la dispersion des revenus dans différents pays, il faut évidemment que les unités monétaires soient uniformisées, mais ce n'est pas tout. Pour prendre un cas extrême, considérons une étude sur les salaires par heure d'une certaine classe d'ouvriers dans deux pays : la Bolivie et la Suisse (on devine déjà qu'on est sur le point de comparer des choses pas comparables). Selon cette étude, les salaires de ces ouvriers ont un écart-type de 0,45 $ (US) en Bolivie et de 4,80 $ en Suisse. Mais il est normal que les salaires suisses, dont la moyenne est de 28 $, varient plus que les salaires boliviens, dont la moyenne est de 3,6 $. Pour comparer les dispersions en tenant compte de ces différences de moyennes, on divise l'écart-type par la moyenne. utilise le *coefficient de variation*, un écart-type *relatif* :

$$\text{cv}(y) = \text{Coefficient de variation de } y = \frac{\text{Écart-type de } y}{\text{Moyenne de } y} = \frac{\sigma_y}{\bar{y}}.$$

Définition *Coefficient de variation*

Soit Y une variable de moyenne $\bar{y}$ et d'écart-type σ_y. Le coefficient de variation cv(Y) est défini par

$$\text{cv}(Y) = \text{Coefficient de variation de } y = \frac{\text{Écart-type de } y}{\text{Moyenne de } y} = \frac{\sigma_y}{\bar{y}} \qquad (1.3.6)$$

En général, si les valeurs de Y sont élevées, on *s'attend* à ce que leur écart-type le soit aussi. Les salaires d'une population d'ouvriers et ceux d'une population de PDG ont des écarts-types difficilement comparables, les chiffres des deux populations n'étant pas du même ordre de grandeur. Le coefficient de variation ramène les écarts-types à des niveaux comparables. Voici, par exemple, quelques données tirées du tableau A.08 sur les poids des hommes et des femmes participant à une certaine recherche sur le cerveau :

	Écart-type	Moyenne	Coefficient de variation
Hommes	8,927 kg	74,19 kg	0,120 ou 12,0 %
Femmes	8,338 kg	61,91 kg	0,135 ou 13,5 %

La dispersion brute est supérieure chez les hommes (écart-type de 8,827 versus 8,338 chez les femmes), mais lorsqu'on tient compte du poids supérieur des hommes (moyenne de 74,19 kg versus 61,91), on trouve que leur dispersion est *relativement* inférieure (coefficient de variation de 12,0 % versus 13,5 %). Dans le tableau suivant, portant sur les *tailles*, l'impression donnée par les écarts-types n'est pas *inversée* par le coefficient de variation, mais elle est atténuée quelque peu: l'écart-type des hommes est de 40 % supérieur à celui des femmes, alors que leur coefficient de variation n'est que de 30 % supérieur.

	Écart-type	Moyenne	Coefficient de variation
Hommes	8.577 cm	180,594 cm	4,7 %
Femmes	6.066 cm	167,529 cm	3,6 %
Ratio Hommes/Femmes	1,41	1,08	1,31

***Exemple* 1.3.1** *Le coefficient de variation ne résout pas tout*

L'exemple suivant est plus complexe. Il présente pour un groupe d'employés, le salaire actuelle et le salaire au moment de l'engagement, ce qui peut dater de plusieurs années :

	Écart-type	Moyenne	Coefficient de variation
Salaire actuel	15 451 $	79873 $	19,3 %
Salaire à l'entrée	11 149 $	17990 $	62,0 %

Une comparaison des écarts-types, sans normalisation, serait fautive, puisque que les salaires à l'entrée, ayant été fixés dans un passé parfois lointain, sont plus faibles. Mais une comparaison des coefficients de variation n'est guère plus utile car le problème n'est pas là. Si les écarts-types ne sont pas comparables, ce n'est pas seulement parce que l'ordre de grandeur n'est pas le même dans les deux séries. C'est aussi parce qu'un salaire de 30000 $ en 2004 n'équivaut pas à un salaire de 30000 $ en 2012 (il équivaut plutôt à 47440 $ compte tenu de l'inflation). C'est ce qui explique la différence démesurée entre les deux coefficients de variation (62,0 % pour les salaires à l'entrée vs 19,3 % pour les salaires en 2012). La dispersion des salaires à l'entrée, même relative, est énorme car elle reflète la variabilité des salaires au cours des années.
[Exprimés en dollars de 2012, les salaires à l'entrée ont une moyenne de 49492 $, un écart-type de 6740 $, et un coefficient de variation de 13,6 %. Donc une plus faible dispersion relative à l'entrée qu'en 2012. Ce qui s'explique par le fait que les salaires en 2012, sont affectés par l'ancienneté, ce qui n'est pas le cas des salaires à l'entrée.] ■

1.4 Quartiles et moustaches

Les figures 1.4.1 et 1.4.2 présentent la distribution du nombre d'heures de travail dans 46 grandes villes du monde. La première le fait de manière classique à l'aide d'un histogramme. La seconde le fait à l'aide d'une *moustache*, une technique employée de plus en plus couramment. Une moustache est construite à partir de certain nombres repères, dont la médiane, et deux quantités appelées des *quartiles*.

Figure 1.4.1

Histogramme du nombre d'heures de travail dans 46 *pays*

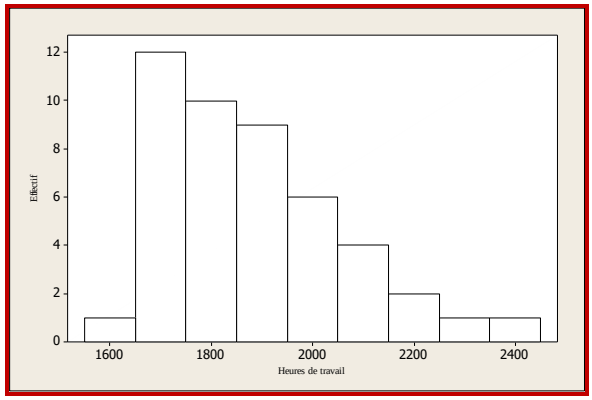
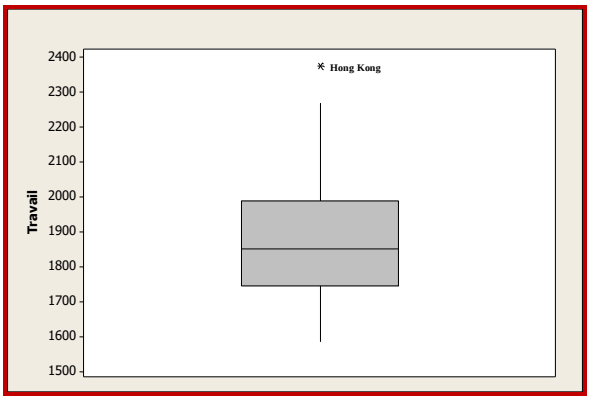
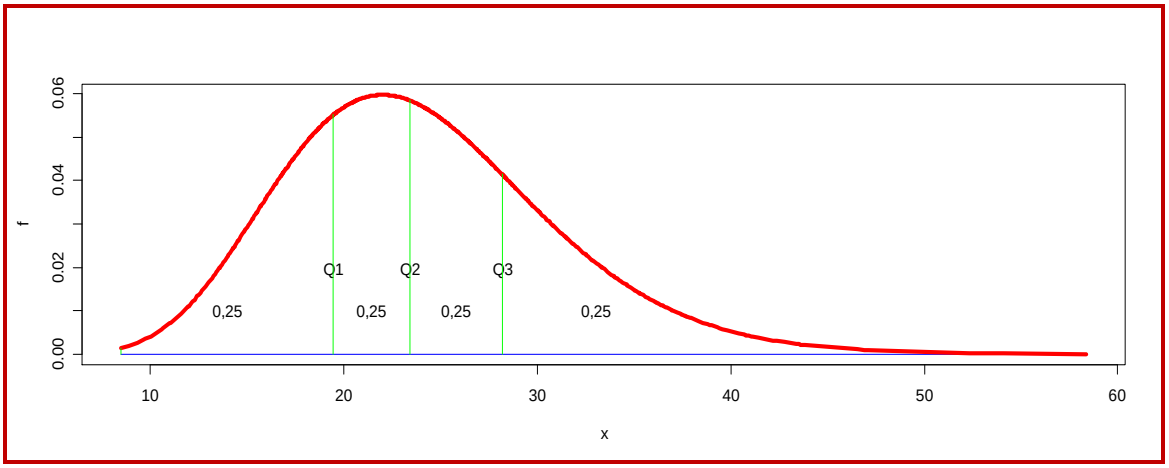


Figure 1.4.2
Moustache représentant la distribution du nombre d'heures de travail dans 46 villes



L'idée d'un quartile est un prolongement de celle d'une médiane: alors que la médiane divise la série en deux parties approximativement égales, les quartiles la divisent en *quatre* parties approximativement égales. Il y a trois quartiles qui les séparent, Q_1, Q_2 et Q_3. On les définira à peu près comme ceci :

- Q_1, le premier quartile : environ 1/4 des données de la série sont inférieures ou égales à Q_1, et environ ¾ des données sont supérieures ou égales à Q_1
- Q_3, le troisième quartile : environ ¾ des données de la série sont inférieures ou égales à Q_3, et environ ¼ des données sont supérieures ou égales à Q_3.
- Q_2, le deuxième quartile est la médiane.



Ces définitions sont quelque peu vagues, car il existe plusieurs façons, pas tout à fait équivalentes, de calculer Q_1 et Q_3 ; et différents logiciels pourraient donner différentes valeurs pour ces quartiles. Mais ces différences ont peu d'impact en pratique car, pour des données suffisamment nombreuses, les différences seront minuscules.

La figure 1.4.2 présente la distribution du nombre d'heures de travail par semaine (pour les ingénieurs) dans 46 villes du monde. Les données sont issues du tableau A.6. Quelques données clé sont présentées ci-dessous :

Minimum	*Limite inférieure normale*	Q_1	*Médiane*	Q_3	*Limite supérieure normale*	*Maximum*
35	35	39,5	40	43	48	51

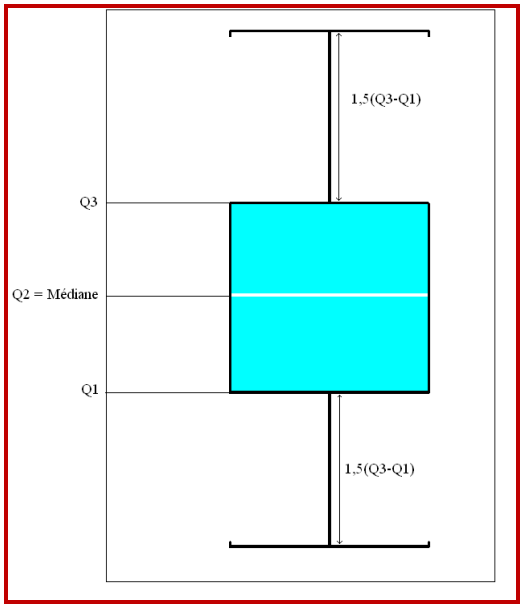
Voici comment, grosso modo, on interprète ce rectangle : la ligne qui traverse le rectangle représente la médiane — 40 — et donc sépare la population en deux parties à peu près égales. Les limites du rectangle représentent à peu près les premier et troisième quartiles et donc englobent environ 50 % de la population. On peut donc dire que la moitié centrale de la population se situe entre 39,5 et 43. Enfin, les limites des deux tiges sont celles dans lesquelles les éléments de la population devraient « normalement » tous se situer. On peut dire que les limites « normales » de la population sont 35 et 48. Finalement, toutes les données qui se situent au-delà des limites des tiges sont marquées individuellement, par un astérisque, un point, ou tout autre symbole qui identifie l'unité.

Les deux tiges sont construites de la façon suivante. Initialement, on trace des lignes imaginaires dont la longueur égale *une fois et demie* la distance entre Q_3 et Q_1. La longueur de ces lignes est donc

$$1{,}5(Q_3 - Q_1) = 5{,}25.$$

Dans une population normale, l'intervalle entre $Q_1 - 1{,}5(Q_3 - Q_1)$ et $Q_3 + 1{,}5(Q_3 - Q_1)$ contient environ 99 % des observations, de sorte qu'il est raisonnable de considérer toute observation à l'extérieur de cet intervalle comme plutôt extrême.

La tige inférieure s'étendrait vers le bas jusqu'à $Q_1 - 5{,}25 = 34{,}25$; et la tige supérieure vers le haut jusqu'à $Q_3 + 5{,}25 = 48{,}25$. Nous ne traçons pas les tiges jusqu'à ces extrêmes, cependant : la tige inférieure ne doit pas descendre plus bas que la donnée la plus petite de l'intervalle. Puisque celle-ci est 35, c'est là qu'on tronque la tige.

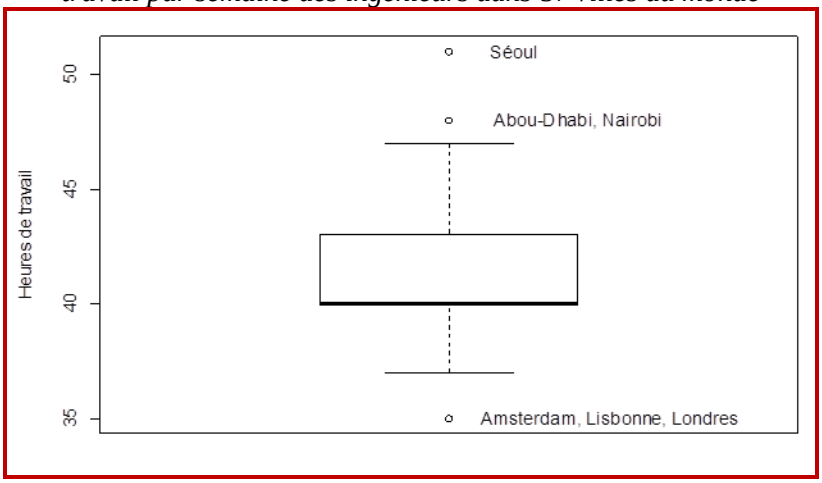


La tige supérieure non plus ne sera pas prolongée jusqu'à la limite de 48,25. Elle sera tronquée à la plus grande valeur inférieure ou égale à ce point, en l'occurrence, 48. La règle générale est celle-ci :

- La limite inférieure de la tige inférieure est
  la plus petite donnée supérieure ou égale à $Q_1 - 1{,}5(Q_3 - Q_1)$;
- la limite supérieure de la tige supérieure est
  la plus grande donnée inférieure ou égale à $Q_3 + 1{,}5(Q_3 - Q_1)$.

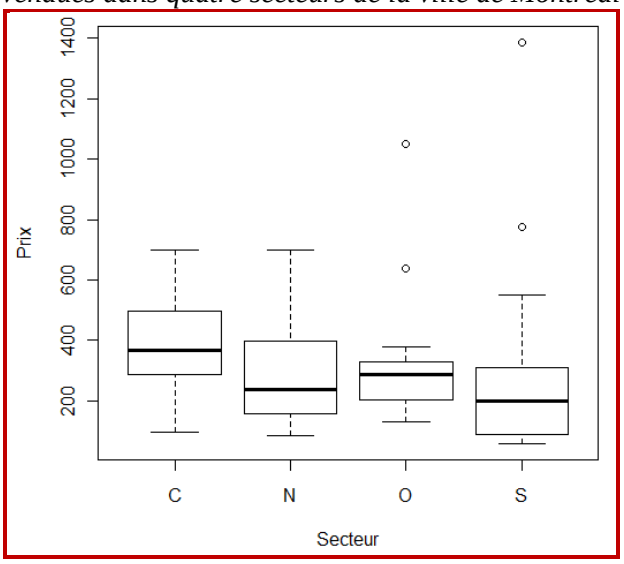
Toute observation qui dépasse les limites est indiquée individuellement. Dans l'exemple, c'est Seoul qui se distingue de par le très grand nombre d'heures de travail par semaine. La présentation par moustaches permet d'attirer l'attention sur ces cas extrêmes. Elle aurait également mis en exergue toute ville dont le nombre d'heure de travail serait exceptionnellement petit, comme dans la figure 1.4.3.

Figure 1.4.3
Moustache représentant la distribution du nombre d'heures de travail par semaine des ingénieurs dans 57 villes du monde



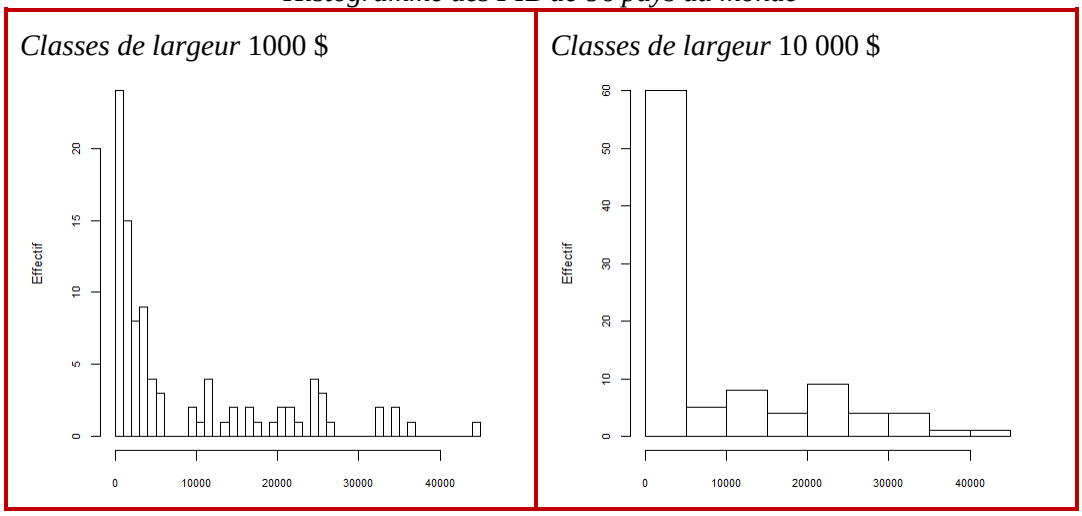
Les moustaches permettent de comparer plusieurs populations. La figure 1.4.4 compare les prix de 102 maisons vendues dans quatre secteurs de Montréal (Centre, Nord, Ouest, Sud). C'est dans le secteur Centre que le prix médian est le plus élevé, et c'est aussi là qu'est la distribution la plus symétrique. C'est dans le secteur Sud que le prix médian est le plus bas, mais on y trouve par ailleurs quelques prix très élevés, une hétérogénéité due à l'existence de quartiers huppés dans ce secteur par ailleurs modeste. Mises à part les quelques données excentriques dans le secteur sud, la dispersion des prix est la même dans chaque secteur sauf dans le secteur Ouest, où les prix sont très concentrés, hormis quelques maisons exceptionnelles.

Figure 1.4.4
Moustaches comparant les prix de 102 *maisons vendues dans quatre secteurs de la ville de Montréal*



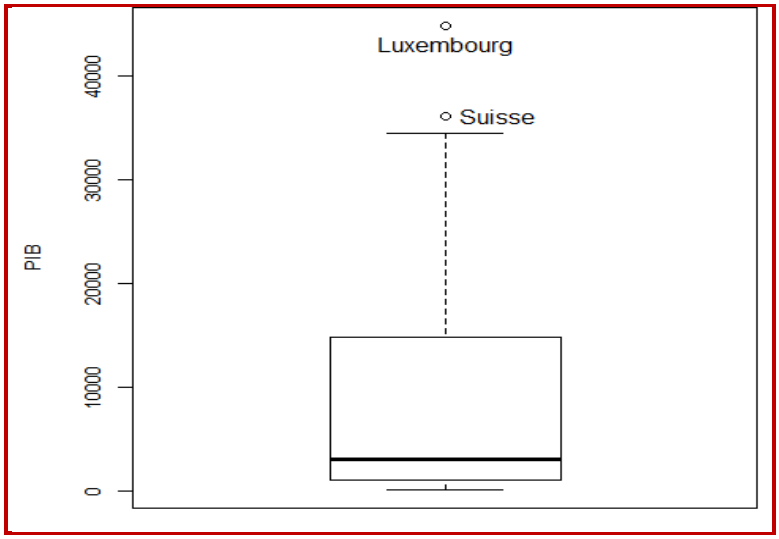
Il existe aussi des distributions qui ne se prêtent tout simplement pas à une représentation par histogramme. C'est le cas des distributions hautement asymétriques où la très grosse majorité des données se situent dans une étroite tranche au bas de l'échelle, comme les PIB des pays du monde : très faibles valeurs pour un grand nombre de pays, suivies de quelques valeurs élevées, dans certains cas démesurées comparées aux premières. La figure 1.4.5 montre clairement que toute tentative de présenter ces données sous forme d'histogramme est vouée à l'échec.

Figure 1.4.5
Histogramme des PIB de 96 *pays du monde*



Une représentation par moustache (Figure 1.4.6) se révèle plus efficace.

Figure 1.4.6
Moustache des PIB de 96 *pays du monde*



Toute observation qui dépasse les limites est indiquée individuellement.

Remarque *D'où vient le* 1,5?
On peut vérifier que dans une population normale, la proportion des valeurs se situant entre -1,5(Q_3 – Q_1) *et* 1,5(Q_3 – Q_1) *est d'environ* 99 %. *Donc toute observation en dehors de cet intervalle est considéré extrême et mérite d'être signalée.*

1.5 Transformations affines et cote Z

Une transformation affine $Y = a + bX$ d'une variable X est une variable Y dont les valeurs sont

$$y_i = a + bx_i \text{ , } i = 1, \dots , n$$

où a et b sont des constantes.

La relation entre la moyenne et la variance de Y et celles de X sont donnés dans le théorème suivant.

Théorème 1.5.1 *Transformation affine*
Soit X une variable de moyenne $\bar{x}$ et de variance σ_x^2 et soit Y une variable définie en fonction de X par $Y = a + bX$, une variable dont chaque valeur x_i est transformée par

$$y_i = a + bx_i \text{ , } i = 1, \dots , n.$$

Alors la moyenne et la variance de Y sont données par

$$\bar{y} = a + b\bar{x} \text{ et } \sigma_y^2 = b^2\sigma_x^2 \text{ .}$$

Démonstration $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(a + bx_i) = \frac{1}{n}\sum_{i=1}^{n} a + \frac{1}{n}\sum_{i=1}^{n} bx_i = \frac{1}{n}na + \frac{b}{n}\sum_{i=1}^{n} x_i = a + b\bar{x}$.

$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n}[a + bx_i - (a + b\bar{x})]^2 = \frac{1}{n}\sum_{i=1}^{n}[bx_i - b\bar{x})]^2 = b^2\frac{1}{n}\sum_{i=1}^{n}[x_i - \bar{x})]^2 = b^2\sigma_x^2$.

***Exemple* 1.5.1** *Transformation affine*

Les températures moyennes X en Arizona durant 6 mois consécutifs ont été rapportées en degrés Fahrenheit :

Valeurs de X : 50°F , 59°F , 68°F , 77°F ,86°F , 95°F

Vous les convertissez en degrés Celsius (Y) :

La transformation est

$$Y = -\frac{160}{9} + \frac{5}{9}X$$

La moyenne et la variance des températures en degrés Fahrenheit sont

$$\bar{x} = 72{,}5 \text{ et } \sigma_X^2 = 236{,}25.$$

par conséquent la moyenne de Y est

$$\bar{y} = a + b\bar{x} = -\frac{160}{9} + \frac{5}{9}\bar{x} = -\frac{160}{9} + \frac{5}{9}72{,}5 = 22{,}5$$

La variance de Y est

$$\sigma_Y^2 = \left(\frac{5}{9}\right)^2 \times \sigma_X^2 = \left(\frac{5}{9}\right)^2 \times 236{,}25 = 72{,}92.$$

L'écart-type de Y est

$$\sigma_y = \left|\frac{5}{9}\right|\sigma_X = \left|\frac{5}{9}\right|15{,}37 = 8{,}54$$ ■

La cote Z

La *cote Z* est une transformation affine particulière, définie par :

$$Z = \frac{X - \bar{x}}{\sigma_X} \tag{1.5.1}$$

La moyenne et l'écart-type de Z sont

$$\bar{z} = 0, \quad \sigma_z = 1 \tag{1.5.2}$$

La cote Z situe un individu par rapport à l'ensemble, ce qui est parfois la seule façon de donner un sens à une donnée numérique. Par exemple, on vous annonce que votre score en un test de dextérité manuelle est de 65. Est-ce bon ou pas? Si vous apprenez que la moyenne de la population (des gens de votre âge et sexe, disons) est de 50, vous savez que vous vous situez à 15 points au-dessus de la moyenne, un écart positif, donc un bon signe. Mais vous ne savez encore pas à quel point vous vous écartez de la moyenne: un écart de 15 points est-il impressionnant? Là aussi, il faut comparer cet écart à quelque chose. Le plus naturel serait de le comparer à l'écart-type de la population, qui est une sorte d' «écart moyen». Supposons que l'écart-type de la population est 5. Alors votre écart à la moyenne est 3 fois l'écart-type. Ce chiffre est la cote Z :

$$Z = \frac{65-50}{5} = 3.$$

La cote Z, donc, exprime la valeur d'une variable par l'écart qui la sépare de la moyenne, avec l'écart-type pour unité de mesure.

Interprétation de la cote Z Quelques repères sont clairs : une cote Z est négative pour une donnée inférieure à la moyenne, positive pour une donnée supérieure à la moyenne, et nulle pour une donnée exactement égale à la moyenne. Outre le signe de Z, comment interpréter sa valeur numérique ? Une cote Z de 3 est-elle importante ou pas ? Voici une première règle générale : une

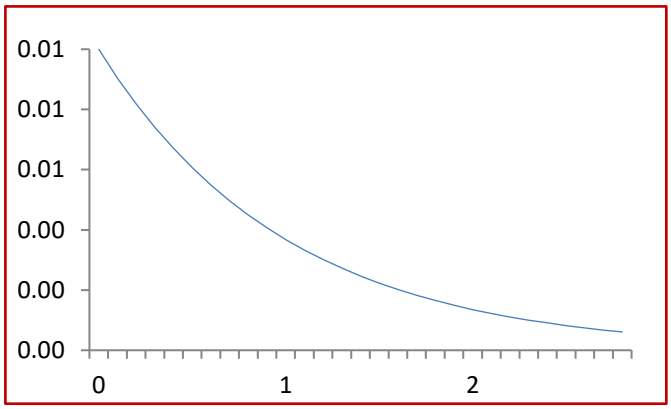
forte proportion de toute population se situe à 2 écarts-types ou moins de la moyenne, c'est-à-dire que pour une forte proportion de la population, on a $|Z| \leq 2$ (ou, ce qui est équivalent, $-2 \leq Z \leq 2$).

Peut-on chiffrer cette proportion ? Oui, à peu près. Supposons, pour commencer, que la population est normale. Si les valeurs de X se distribuent selon une loi normale, leurs cotes Z se distribuent selon une loi normale centrée-réduite. On peut alors affirmer ceci (voir l'illustration, figure 1.5.3)

$$P(|Z| \leq 1) \approx 0{,}68;\ \ P(|Z| \leq 2) \approx 0{,}95;\ \ P(|Z| \leq 3) = 0{,}997.$$

La notation $P(|Z| \leq a)$ désigne la proportion de la population dont les valeurs Z satisfont la condition $|Z| \leq a$.

Remarque *Les populations ne sont pas toutes normales, bien sûr. On peut néanmoins considérer ces pourcentages comme indices approximatifs. Ce sont des approximations adéquates dans un bon nombre de populations. À titre de comparaison, considérons une population de forme exponentielle, comme ceci :*



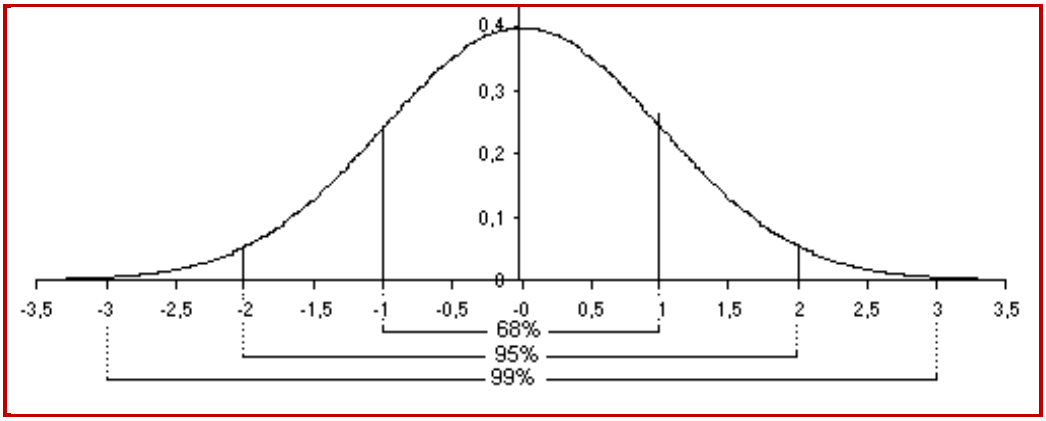
Le tableau suivant compare une population normale à une population exponentielle — très éloignée d'une population normale — par rapport au pourcentage de données dont la cote Z, en valeur absolue, est inférieure à un nombre a :

Proportion de la population pour laquelle $|Z| \leq a$ selon la population

a	1	1,5	2	2,5	3	3,5	4	4,5
Population Exponentielle	86,5%	91,8%	95,0%	97,0%	98,2%	98,9%	99,3%	99,6%
Population Normale	68,3%	86,6%	95,4%	98,8%	99,7%	100,0%	100,0%	100,0%

Les différences, on le voit, sont pour la plupart négligeables. ■

Figure 1.5.3
Loi normale



Lorsqu'on ne connaît pas la distribution de la population, et qu'on ne peut pas supposer la normalité, on ne peut pas calculer P($|Z| \geq 3$) ou P($|Z| < 3$). On peut, cependant, déterminer une borne supérieure P($|Z| \geq a$), pour un nombre positif a arbitraire, grâce à un théorème appelé *Inégalité de Tchebychev*.

Théorème 1.5.2 *Inégalité de Tchebychev*
Soit X une variable et a un nombre positif. Alors

$$P(|Z| \geq a) \leq \frac{1}{a^2}$$

C'est ce qui permet d'affirmer, par exemple, qu'il y a peu d'observations pour lesquelles $|Z| \geq 3$, car d'après le théorème, $P(|Z| \geq 3) \leq \frac{1}{3^2} = \frac{1}{9}$ (et $P(|Z| < 3) \geq 1 - \frac{1}{9} = \frac{8}{9}$).

1.6 Corrélation et droite des moindres carrés

Les études statistiques ne portent pas toujours sur une seule variable. Au contraire, elles portent souvent sur plusieurs variables à la fois, et en particulier, sur les *relations* entre elles. Nous présentons ici des techniques permettant d'examiner la relation entre deux variables, X et Y, dont les valeurs sont appariées, c'est-à-dire que, pour un individu i donné, les valeurs x_i et y_i sont des mesures prises sur un même individu. Il convient donc de parler d'un couple [X ; Y] dont les valeurs sont [x_1 ; y_1]; [x_2 ; y_2]; …; [x_n ; y_n].

Ces couples peuvent être représentés graphiquement par un *nuage de points*. Considérons, par exemple, les données concernant le salaire (Y, en milliers de dollars) et l'ancienneté (X) d'un groupe de 200 professeurs. Ces données prennent la forme de deux séries de 200 chiffres dont voici les quelques premiers :

Ancienneté (x)	Salaire en 2012 (y)	Ancienneté (x)	Salaire en 2012 (y)	Ancienneté (x)	Salaire en 2012 (y)	Ancienneté (x)	Salaire en 2012 (y)
17	69 526	22	101 508	5	68 334	22	80 650
28	85 418	13	97 734	11	80 650	5	58 402
4	51 647	17	89 192	8	69 924	20	87 007

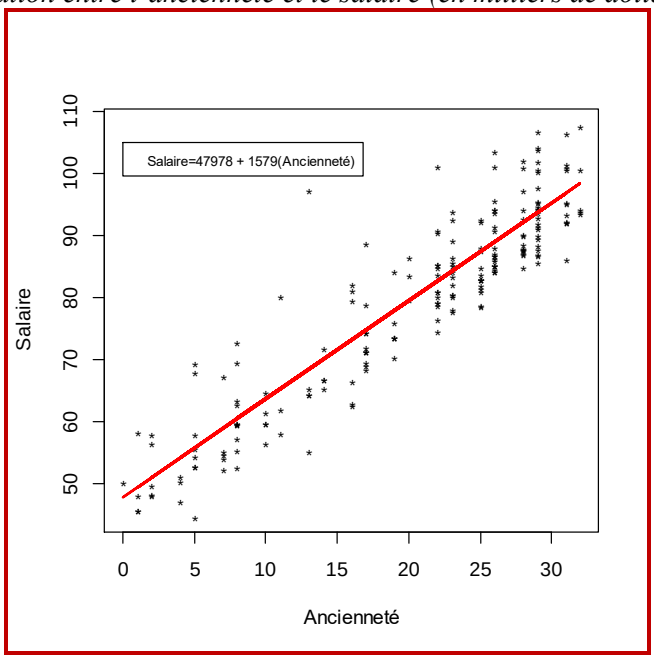
Le nuage de points (figure 1.6.1) donne une première impression nette et visuelle de la nature et de la force de la dépendance entre les deux variables. Il montre clairement qu'il y a une relation entre les deux variables, et que cette relation est *croissante* : plus l'employé est ancien, plus il gagne. De plus, cette croissance semble *linéaire* : on pourrait tracer une droite autour de laquelle les points se concentrent.

Il reste à quantifier ces observations.

Dans cette section, nous allons

- Pour commencer, définir une façon de mesurer le degré de dépendance, car certaines relations sont très fortes, d'autres moins; et
- S'il se trouve qu'il y a une relation et si, en plus, il existe une droite qui exprime — au moins approximativement — la relation entre les variables, nous voudrons en déterminer l'équation.

Figure 1.6.1
Nuage de points
Relation entre l'ancienneté et le salaire (en milliers de dollars)



Mesure de dépendance

Une mesure de dépendance, la *covariance*, désignée par Cov(X ; Y) ou σ_{xy}, est définie par

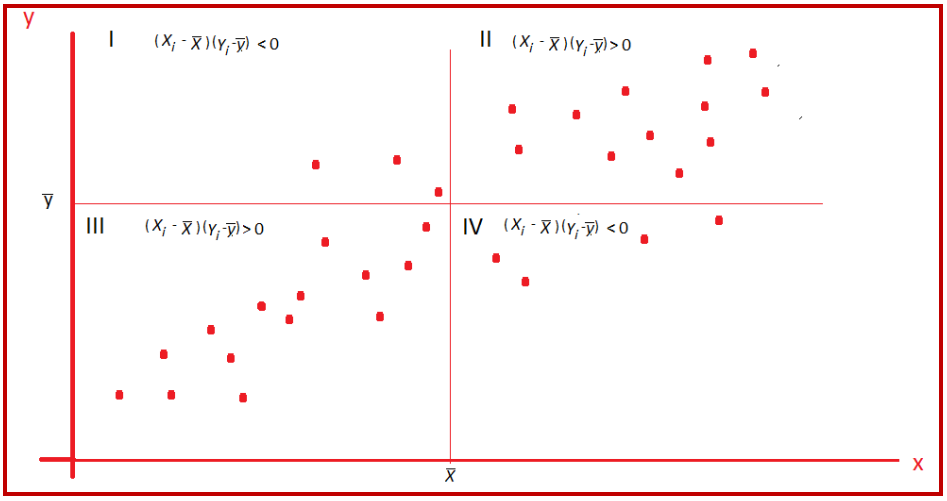
$$\text{Cov}(X\,;\,Y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} \qquad (1.6.1)$$

(Voir l'exemple 1.6.1 ci-dessous pour une illustration du calcul.)

Une forte dépendance entre X et Y est caractérisée par une covariance élevée en valeur absolue. Et deux variables indépendantes ont une covariance nulle. La figure 1.6.2 montre en quoi la covariance est une mesure de dépendance. Supposons qu'il y a une dépendance positive entre X et Y. On verra alors plusieurs points dans les quadrants II et III et peu de points dans les quadrants I et IV. Dans ces quadrants, les produits $(x_i - \bar{x})(y_i - \bar{y})$ sont positifs et la valeur de leur somme

ne sera compensée que par les peu nombreux produits $(x_i - \bar{x})(y_i - \bar{y})$ négatifs dans les quadrants I et IV. Si la relation est négative, ce seront les produits $(x_i - \bar{x})(y_i - \bar{y})$ (négatifs) dans les quadrants I et IV qui dominent.

Figure 1.6.2
Nuage de points
Relation entre l'ancienneté et le salaire (en milliers de dollars)



Cet indice est imparfait. Pourquoi ? Parce sa valeur *dépend de l'unité de mesure*. La covariance entre la taille et le poids de *n* personnes, par exemple, changera selon que la taille est mesurée en mètres ou en pieds; et selon que le poids est mesuré en kilogrammes ou en livres. Le *Coefficient de corrélation* vise à pallier cette difficulté

Définition *Coefficient de corrélation*
Le *Coefficient de corrélation* entre deux variables, *X* et *Y*, est défini par

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1.6.2}$$

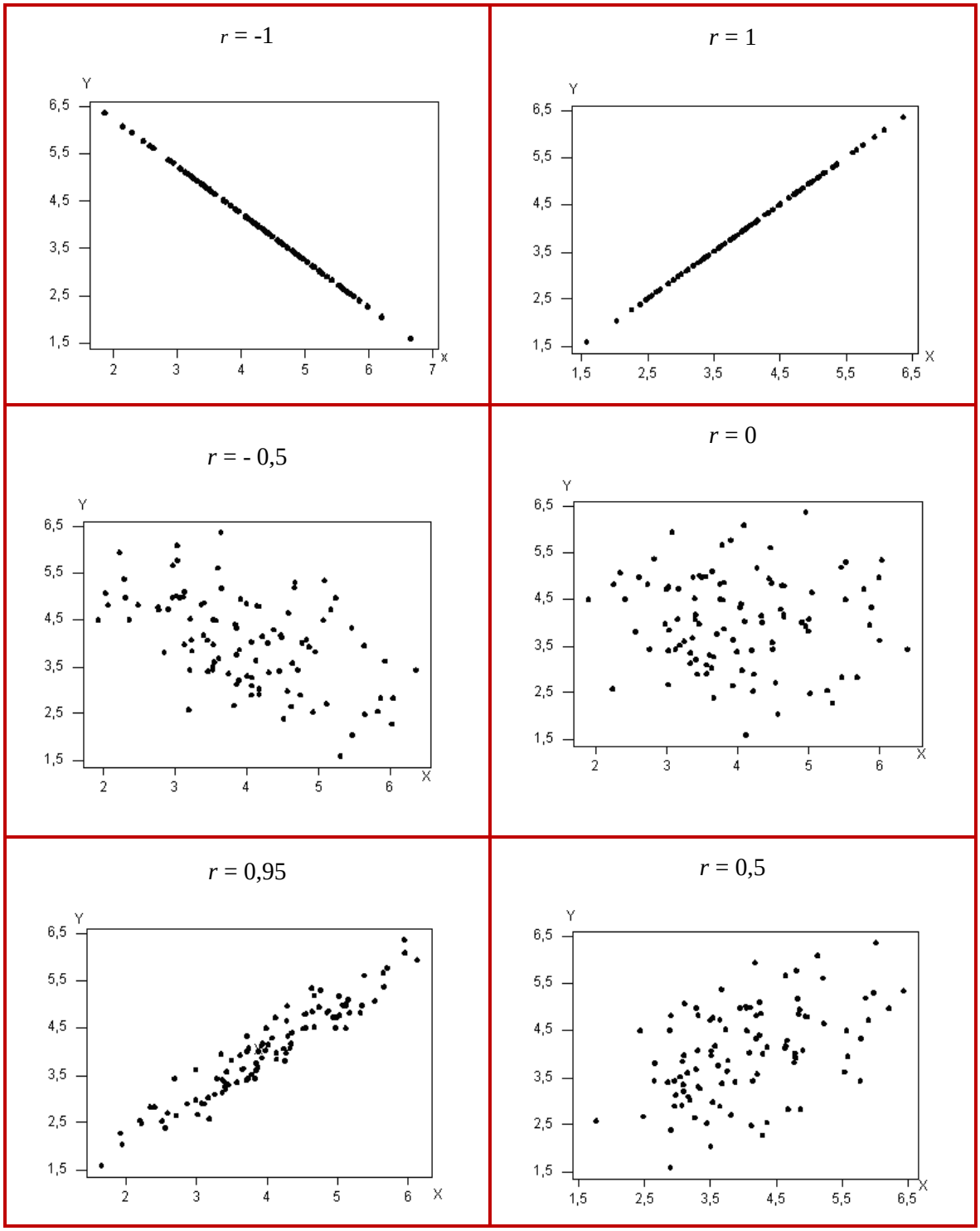
Essentiellement, le coefficient de corrélation permet de rendre la covariance indépendante des unités de mesure en remplaçant les données x_i et y_i par leur cote *Z*, $z_{x_i} = \frac{x_i - \bar{x}}{\sigma_x}$ et $z_{y_i} = \frac{y_i - \bar{y}}{\sigma_y}$.

Ce qui en résulte est le coefficient de corrélation, car

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \text{Cov}\left(\frac{X - \bar{x}}{\sigma_x}; \frac{Y - \bar{y}}{\sigma_y}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \tag{1.6.3}$$

Figure 1.6.3

Coefficient de corrélation - Quelques exemples



Propriétés du coefficient de corrélation :

1) $-1 \leq r \leq 1$: *le coefficient de corrélation se situe toujours entre* -1 *et* 1 (*inclusivement*).

2) $|r| = 1$ *si et seulement si les points du nuage se situent tous sur une même droite.*

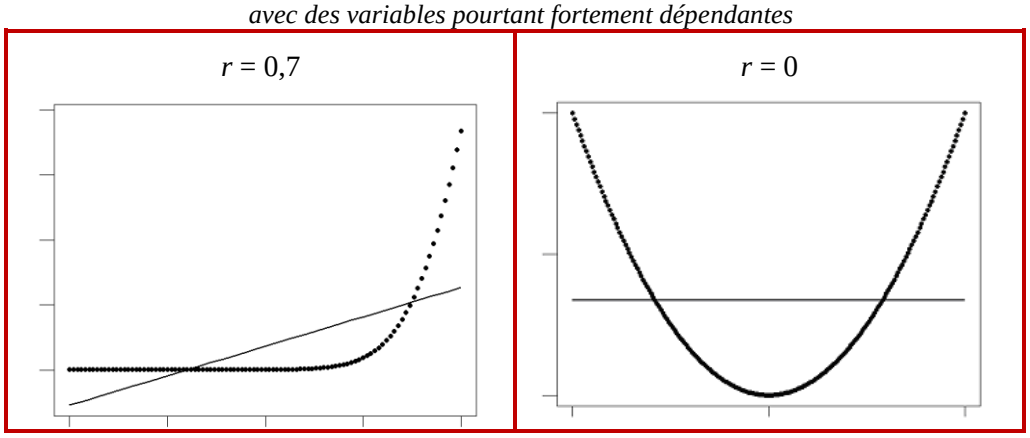
3) *Le coefficient de corrélation mesure le degré d'alignement des points du nuage le long d'une* droite. *Si la droite est de pente positive, r est positif ; si la droite est de pente négative, r est négatif.*

4) *Le coefficient de corrélation est élevé* (*en valeur absolue*) *dans la mesure où les points du nuage se rapprochent d'une droite non horizontale. Lorsque les points se placent tous sur une même droite, la corrélation est parfaite. Dans ce cas, r* = 1 *si la droite est de pente positive et r* = -1 *si la droite est de pente négative.*

5) *Lorsque les variables sont indépendantes, le coefficient de corrélation est nul : r* = 0. *La figure* 1.6.3 *montre l'allure du nuage de points correspondant à quelques valeurs de r.*

6) *La réciproque de* 5) *n'est pas vraie : le coefficient de corrélation peut être faible alors que la relation est en fait très forte. C'est que r est une mesure de dépendance* linéaire. *Une dépendance forte mais non linéaire peut avoir un coefficient de corrélation relativement faible ou même nul. Voir la figure* 1.6.4. ■

Figure 1.6.4
Coefficient de corrélation relativement faible ou nul avec des variables pourtant fortement dépendantes



Remarque *Autres versions des formules*

Il existe une autre version des formules de σ_{xy} *et de* σ_x^2 *qui facilitent les calculs manuels. Les voici :*

$$\sigma_x^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n} = \overline{x^2} - \bar{x}^2 \text{ ; et } \sigma_{xy} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\,\bar{y}}{n} = \overline{xy} - \bar{x}\,\bar{y} \qquad (1.6.4)$$

Droite des moindres carrés

Lorsque le nuage de points suggère l'existence d'une relation linéaire entre deux variables, il est utile de déterminer la droite, $y = b_0 + b_1 x$, disons, qui exprime cette relation. C'est une droite autour de laquelle les points se concentrent. La droite qu'on choisira devra passer le plus près possible des points du nuage. Pour préciser cette notion, nous commençons par définir une mesure de la distance entre le nuage et la droite. Celle que nous adoptons est une quantité Q définie par :

$$Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \qquad (1.6.5)$$

où $\hat{y}_i = b_0 + b_1 x_i$. Q est donc la somme des carrés des distances verticales entre les points du nuage et la droite.

La *droite des moindres carrés* est celle qui minimise *Q*. On trouve par les méthodes du calcul différentiel que les valeurs de b_0 et b_1 qui minimisent *Q* sont

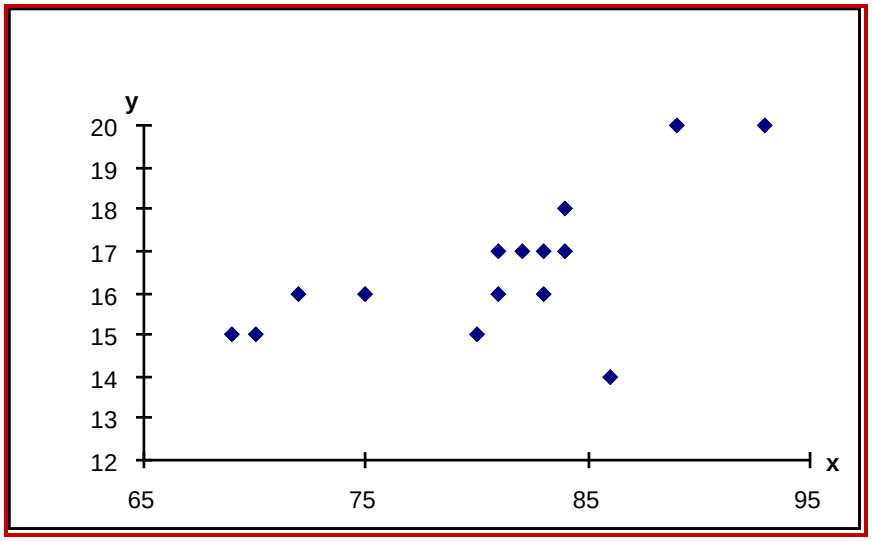
$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad \text{et} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Remarque *Il est intéressant de noter que la droite passe par le point* $(\bar{x}\ ; \bar{y})$*, ce qui signifie que pour une valeur moyenne de x, on prévoit une valeur de y… moyenne aussi.* ■

***Exemple* 1.6.1** Y a-t-il une relation entre la température (*x*) et le nombre de cris du criquet (*Y*) ? Les données suivantes ont été prélevées pour répondre à cette question.

x	89	72	93	84	81	75	70	82	69	83	80	83	81	84	86
y	20	16	20	18	17	16	15	17	15	16	15	17	16	17	14

Solution Le graphique montre qu'il existe effectivement une certaine dépendance :



Les sommes, sommes de carrés, et somme de produits sont :

$$\sum x_i = 1212\ ;\ \sum y_i = 249\ ;\ \sum x_i^2 = 98572\ ;\ \sum y_i^2 = 4175\ ;\ \sum x_i y_i = 20227.$$

Le coefficient de corrélation est

$$r = \frac{20227 - (1212)(249)/15}{\sqrt{98572 - (1212)^2/15}\sqrt{4175 - (249)^2/15}} = 0,6594$$

On calcule la droite de régression :

$$b_1 = \frac{20227 - (1212)(249)/15}{98572 - (1212)^2/15} = 0,16780822\ ;$$

$$b_0 = \left(\frac{249}{15}\right) - 0,16780822\left(\frac{1212}{15}\right) = 3,0411$$

La droite de régression est donc

$$y = 3{,}0411 + 0{,}1678\, x$$ ■

RÉSUMÉ

1 Une *variable* fait correspondre une *valeur* à chacune des *unités statistiques* de la *population*. Une variable est dite *quantitative* si ses valeurs sont des nombres représentant des quantités ; autrement elle est dite *qualitative*. Une variable est dite *discrète* si l'ensemble de ses valeurs est fini ou dénombrable (c'est-à-dire, s'il existe une bijection entre l'ensemble des valeurs et les nombres naturels). Autrement elle est dite *continue*.

2 Une *distribution* est une fonction qui fait correspondre à chaque valeur d'une variable un *effectif* ou une *fréquence*. Une *distribution cumulative* fait correspondre à chaque valeur d'une variable l'effectif cumulé ou la fréquence cumulée. Lorsque les données sont groupées, on fait correspondre un effectif ou une fréquence à des intervalles.

3 Une distribution est généralement représentée graphiquement par un *digramme à bâtons*, un *histogramme*, un *polygone des fréquences* ou une *moustache*.

4 La *moyenne arithmétique* est définie par $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ pour une série de données et par $\bar{x} = \frac{1}{n}\sum_{i=1}^{p} x_i n_i = \sum_{i=1}^{p} x_i f_i$ pour une distribution. Dans le cas d'une distribution, les x_i représentent soit les *points-milieux*, soit les valeurs elles-mêmes, selon que les valeurs sont groupées ou non.

5 La *médiane* d'une série de données est le centre de la série lorsque les données sont placées en ordre ; lorsque le nombre de données est pair, la médiane est la moyenne arithmétique des données centrales. Le *mode* est la donnée la plus fréquence. Lorsque les données sont groupées, on parle plutôt de *classe modale*.

6 La *variance* est définie par $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ pour une série de données et par $\sigma^2 = \sum_{i=1}^{p}(x_i - \bar{x})^2 f_i$ pour une distribution. L'*écart-type* est la racine carrée de la variance.

Formules alternatives : $\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$ pour une série de données, et $\sigma^2 = \sum_{i=1}^{p} x_i^2 f_i - \bar{x}^2$ pour une distribution.

7 Si $Y = a + bX$, alors $\bar{y} = a + b\,\bar{x}$ et $\sigma_y = |b|\sigma_x$.

La *cote Z*, définie par $Z = \frac{X - \bar{x}}{\sigma}$, est de moyenne nulle et de variance 1.

8 Le *coefficient de corrélation r* est une mesure de dépendance linéaire définie par

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \text{ où } \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n}$$

9 Les coefficients de la droite des moindres carrés sont :

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \quad \text{et} \quad b_0 = \bar{y} - b_1\bar{x}$$

EXERCICES

1.1 Déterminer la moyenne et l'écart-type des données suivantes a) d'abord en utilisant la série entière telle quelle, ensuite b) après les avoir disposées sous forme de distribution.

Nombre de pièces dans un échantillon de 61 *logements*

2	2	2	2	2	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4	4	5	5	5	5	5	5	5	5	5	5	6	6	6
6	6	6	6	6	7	7	7	7	7	8	8	8	8	9
9														

1.2 Déterminez la moyenne arithmétique et l'écart-type de la distribution suivante :

y	3	6	8	9	11	Total
Fréquence	0,2	0,4	0,2	0,1	0,1	1

1.3 Laquelle des deux séries suivantes semble plus dispersée ? Répondre sans calculer.

A :	50	59	60	61	70
B :	18	19	20	21	22

1.4 Laquelle des deux séries suivantes semble plus dispersée ? Répondre sans calculer.

A :	30	40	50	60	70
B :	10	29	30	31	50

1.5 Laquelle des deux variables suivantes vous semble la plus dispersée ? Répondre sans calculer :

x	1	2	3	4	5	*Total*
Effectif	6	2	4	2	6	20

y	4	5	6	7	8	*Total*
Effectif	2	4	8	4	2	20

1.6 La moyenne et la variance d'une série de températures quotidiennes, en degrés Celsius, sont respectivement 18 et 25. Déterminez la moyenne, la variance et l'écart-type de la même série, exprimée en degrés Fahrenheit.

1.7 Voici une série de 6 températures enregistrées en Arizona, exprimées en degrés Fahrenheit (*X*)

Valeurs de X : 50°F, 59°F, 68°F, 77°F, 86°F, 95°F

a) Déterminer la moyenne $\bar{x}$ et l'écart-type σ_X de *X*.

b) Convertir les températures en degrés Celsius, c'est définir un nouvelle variable, *Y*, fonction de *X*, définie par $Y = -\frac{160}{9} + \frac{5}{9}X$. Vérifiez que les valeurs de *Y* sont :

Valeurs de *Y* : 10°F, 15°F, 20°F, 25°F, 30°F, 35°C

c) Déterminer la moyenne $\bar{y}$ et l'écart-type σ_Y de *Y*, d'abord directement à partir des valeurs de *Y*, ensuite en utilisant les règles $\bar{y} = a + b\bar{x}$ et $\sigma_Y = |b|\sigma_X$ lorsque *Y* est définie par $Y = a + bX$.

d) Convertissez toutes les valeurs de *X* et toutes les valeurs de *Y* en cotes *Z*. Vous devriez constater que les deux séries sont identiques. Calculez la moyenne $\bar{z}$ et l'écart-type σ_Z des cotes *Z*. Vous devriez constater que $\bar{z} = 0$ et $\sigma_Z = 1$. De quelle propriété générale ces observations découlent-elles ?

1.8 Démontrez les propriétés $\bar{y} = a + b\bar{x}$ et $\sigma_y = |b|\sigma_x$, où *X* est une variable et $Y = a + bX$. Déduisez que les cotes *Z* sont de moyenne nulle et d'écart-type 1.

1.9 Un médecin vous dit que votre pression intraoculaire est de 23. Pour une population de 100 000 personnes de votre âge, la pression moyenne est de 17 avec un écart-type de 2,1. Votre pression est-elle excessive ?

1.10 Soit X le revenu des corporations multinationales du Canada ; et soit Y le revenu annuel des petites et moyennes entreprises du Canada. D'après vous, l'écart-type de X est-il supérieur ou inférieur à celui de Y ? Discuter.

1.11 Soit A la série des 365 températures quotidiennes à Montréal (pour une année donnée) et B la série des 365 températures quotidiennes à Miami (même année.) D'après vous, laquelle des deux séries a la plus grande variance ?

1.12 Considérons les variables X et Y, où X représente la proportion quotidienne de garçons parmi les nouveau-nés d'un petit hôpital et Y la proportion quotidienne parmi tous les nouveau-nés canadiens. D'après vous, laquelle des deux variables a le plus grand écart-type ? Discutez.

1.13 Voici une série de valeurs accouplées x et y :

x	4	6	8	12	15
y	5	12	9	12	22

Déterminer les moyennes de X et de Y ; les écarts-types de X et Y ; la covariance entre X et Y ; les coefficients b_1 et b_0 de la droite de régression de Y sur X ; et le coefficient de corrélation r.

1.14 Une usine fabrique des toiles métalliques pour des usines de pâtes et papier. Afin de répartir son personnel, le gérant aimerait prévoir le temps, T, requis pour la finition des toiles. Ce temps pourrait être lié, entre autres variables, à la surface de la toile, S. On a obtenu les données du tableau 1 :

a) Faire un graphique des données. Tracer la droite de régression. Le modèle est-il raisonnable ?

b) Quelle variable doit-on utiliser comme variable dépendante ? (Justifier ce choix).

c) Déterminer l'équation de régression correspondante et le coefficient de corrélation

d) Quel est le temps moyen de finition pour une toile de 20 m^2 ?

Tableau 1

Temps de finition d'une toile (T) et surface de la toile (S)

i	*T*	*S*	*i*	*T*	*S*
1	5,50	9,30	9	6,50	15,80
2	5,90	13,50	10	6,50	14,90
3	5,80	11,10	11	7,10	18,60
4	6,30	14,90	12	7,00	15,80
5	7,00	16,70	13	6,90	16,70
6	7,50	23,20	14	6,80	15,80
7	5,50	11,10	15	6,60	16,70
8	7,20	20,40			

1.15 Un professeur de secondaire est responsable de l'enseignement de l'algèbre. Au début de l'année, il fait passer à 20 de ses étudiants un petit test mesurant les habiletés arithmétiques (H) de ses étudiants. À la fin du premier semestre, il examine les résultats (F) de ses étudiants à l'examen d'algèbre. Les résultats sont présentés au tableau 2 :

a) Faire un graphique des données. Tracer la droite de régression. Le modèle est-il raisonnable ?

b) Quelle variable doit-on utiliser comme variable dépendante ? (Justifier ce choix).

c) Déterminer l'équation de régression correspondante et calculer le coefficient de corrélation entre les deux variables

d) Quelle note à l'examen d'algèbre aurait un étudiant dont la note au premier aurait été 25 ?

Tableau 2

Habileté mathématique (H) et résultat à un examen d'algèbre (F)

i	*F*	*H*	*i*	*F*	***H***
1	36	9	11	59	26
2	23	10	12	58	28
3	22	13	13	72	30
4	36	15	14	87	31
5	49	16	15	86	32
6	32	18	16	79	33
7	44	20	17	74	34
8	52	22	18	78	36
9	51	23	19	99	38
10	83	24	20	85	40

1.16 Les données suivantes présentent le nombre de bactéries *N* encore vivantes après avoir été exposées à des rayons *X* pendant un temps de durée *t.*

N	355	211	197	166	142	106	104	60	56	38	36	32	21	19	15
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

a) Déterminer la droite des moindres carrés pour exprimer *N* comme fonction linéaire de *t.* Déterminer le coefficient de corrélation.

b) Supposons que le phénomène soit assez bien connu pour savoir que la relation entre *N* et *t* est de la forme $N = b_0 e^{-b_1 t}$. De là, on peut conclure que $Y = \ln N$ est une fonction linéaire de *t*. Donc remplacez *N* par *Y* et décidez si l'ajustement ici est meilleur.

1.17 [*Données du tableau* A.1] Le tableau A.1 présente des données sur les professeurs d'une certaine université.

a) Représentez par des moyens graphiques la distribution de la variable « Département ».

b) On compare ici les salaires à l'entrée et les salaires en 2012. Vous auriez intérêt à exprimer les salaires en milliers, de façon à ne pas être encombré de gros chiffres.

(i) Faites deux polygones de fréquences représentant la distribution des salaires à l'entrée et la distribution des salaires en 2012. Faites un commentaire sur les différences entre les deux distributions.

(ii) Déterminez la moyenne, la médiane et l'écart-type des deux distributions : ces deux mesures confirment-elles les impressions visuelles ?

(iii) Les positions relatives des médianes par rapport aux moyennes arithmétiques ne sont pas les mêmes dans les deux distributions. Expliquez ce fait en termes des caractéristiques visuelles des histogrammes. Suggérez aussi une explication dans les termes du contexte.

(iv) La dispersion des salaires en 2012 est bien plus importante qu'à l'entrée. Mais les écarts-types sont-ils réellement comparables ? Une mesure de dispersion *relative,* appelée *coefficient de variation* et définie par CV = $\frac{\sigma}{\bar{x}}$, peut aider à répondre à cette question. Calculez le coefficient de variation des salaires en 1991 et à l'entrée. Pouvez-vous expliquer le fait que maintenant c'est le salaire *à l'entrée* qui est plus dispersé ?

c) Examinez la relation entre le salaire en 2012 et l'ancienneté : faites un graphique, déterminez la droite de régression s'il y a lieu, calculez le coefficient de corrélation.

d) On s'intéresse au lien entre le salaire à la date d'engagement et l'expérience.

(i) Déterminez un nuage de points permettant de voir s'il y a une relation entre le salaire à la date d'engagement et l'expérience.

(ii) Déterminez la droite des moindres carrés. D'après votre droite, qu'est-ce qu'une année de plus rapporte en salaire ?

(iii) Déterminez le coefficient de corrélation entre la date d'engagement et le salaire à l'entrée ; et entre la date d'engagement et le salaire en 2012.

e) On s'intéresse ici à la relation entre la date d'entrée et le salaire à l'entrée. (Vous pourriez choisir de transformer les données sur les années en soustrayant un même nombre (1979, par exemple) de chaque donnée).
   (i) Faites un nuage de points pour montrer la relation entre le salaire à l'entrée et la date d'entrée.
   (ii) Utilisez la droite des moindres carrés en (i) pour prédire le salaire (à la date d'engagement) pour chaque sujet. Construisez une colonne des différences entre les salaires et la prédiction du salaire. Calculez l'écart-type de ces différences. Comment cet écart-type se compare-t-il à l'écart-type des salaires à la date d'engagement ? Pouvez-vous expliquez pourquoi il y a une aussi grande différence ?
   (iii) Le nuage de points obtenu en (i) montre clairement que la relation n'est pas linéaire. Si on suppose que les salaires se sont accrus à un *taux* constant ce dernières décennies, alors la relation est en fait exponentielle : $y = \alpha e^{\beta x}$. Auquel cas, $\ln y = \ln \alpha + \beta x$. Examinez à l'aide d'un nuage de points la relation entre le logarithme du salaire et la date d'engagement.
   (iv) Bien que le graphique en (iii) demeure quelque peu convexe, déterminez la droite des moindres carrés qui lie le logarithme du salaire à l'année d'engagement. Estimez le salaire moyen d'un professeur engagé en 2000.
   (v) Montrez que le pourcentage d'accroissement annuel dans une relation de la forme $y = \alpha e^{\beta x}$ est de $100(e^{\beta} - 1)$ %. Estimez le pourcentage annuel d'accroissement (t) à partir des résultats en (iv).

f) Revenons au salaire à l'entrée et l'expérience. Utilisez le taux d'accroissement t obtenu en *e*) (iv) pour ajuster les salaires y à l'entrée, c'est-à-dire, pour les exprimer en dollars de 2012. Il suffit de multiplier y par $(1+t)^k$, où k est la différence entre 2012 et l'année d'engagement. Maintenant calculez l'écart-type et le coefficient de variation des salaires ajustés. Faites un commentaire sur les différences entre ces mesures et celles obtenues pour les salaires de l'année 2012. Est-ce que les écarts entre professeurs s'accentuent ?

1.18 [*Données du tableau* A.2] Le tableau A.2 présente quelques données sur 43 maisons vendues. Présentez des tableaux, des graphiques, ou des mesures descriptives qui permettent de confirmer ou d'infirmer les propositions suivantes (pour les besoins de cet exercice, une « vieille » maison est une maison de plus de 10 ans) :

a) Les vieilles maisons ont moins souvent un sous-sol. Répondez de deux façons : *i*) en considérant l'âge comme variable quantitative ; et ii) en la considérant comme variables dichotomique : 1 = vieille, 0 = pas vielle.

b) Les maisons qui ont deux salles de bains ou plus coûtent en moyenne 10 000 $ de plus que celles qui en ont moins de deux.

c) Le fait d'avoir deux places de garage ajoute plus à la valeur d'une vieille maison qu'à celle d'une moins vieille.

d) En général, plus une maison est vieille, moins elle coûte.

e) Le prix des maisons décroît avec l'âge, mais c'est surtout parce que les vieilles maisons ont moins souvent deux salles de bains.

1.19 [*Données du tableau* A.4] Le tableau A.4 en annexe présente des données sur une expérience dont l'objet est de comparer trois méthodes d'enseignement chez des enfants. L'objectif de cet enseignement est de parfaire la compréhension de texte. Chaque sujet a composé deux pré-tests (A1, A2) avant la période d'apprentissage et 3 post-tests (B1, B2, B3) après. Comparez la méthode 1 à la méthode 2 de deux façons :

a) Utilisez la moyenne de A1 et A2 comme mesure de compréhension *avant* la période d'apprentissage, et la moyenne de B1 et B2 comme mesure de compréhension après la période d'apprentissage. Comparez les deux groupes par rapport à la différence entre ces deux moyennes (vous supposerez que les scores aux tests A1, A2, B1, B2 sont comparables, de sorte qu'il est raisonnable de les additionner, ou de soustraire l'un de l'autre). À première vue, quelle est la méthode la plus prometteuse ?

b) Comparez les méthodes 1 et 2 en n'employant cette fois-ci que la variable B3 comme mesure de compréhension après la période d'apprentissage (le test B3 est de nature différente des autres ; il n'a été donné *qu'après* la période d'apprentissage).

c) En *b*), on compare les groupes par rapport à un post-test seulement. Est-ce valable ? Si oui, y a-t-il un avantage à procéder de cette façon, ou est-ce préférable d'utiliser la méthode pré-test/post-test ?

1.20 [*Données du tableau* A.5] Le tableau A.5 en annexe présente des données sur la température (en degrés Fahrenheit) de 113 sujets (prises par les sujets eux-mêmes).

a) Il est bon de vérifier de façon empirique ce qui se démontre formellement. Convertissez les températures dans ce tableau (x) en degrés Celsius (y) et utilisez ces données pour vérifier les propriétés $\bar{y} = a + b\,\bar{x}$, $\sigma_y^2 = b^2\sigma_x^2$, et $\sigma_y = |b|\sigma_x$

b) Dans le même esprit qu'en a), utilisez les données sur la température pour montrer que les cotes Z sont de moyenne nulle et d'écart-type égal à 1.

c) On utilise souvent en statistique la *loi normale*, une courbe symétrique en forme de cloche : plusieurs variables dans la nature semblent suivent une loi normale. Faites un histogramme ou un polygone des fréquences de la distribution des températures. Cette distribution vous semble-t-elle normale ? (*Suggestion* : Employez, cette fois-ci, des classes de largeur 0,5, pour réduire les variations aléatoires).

d) Quel intervalle de températures considérez-vous normal ? À partir de quelle valeur diriez-vous qu'une température est excessive ? Considérez toute valeur éloignée de plus de 2,5 écarts-types de la moyenne comme étant excessive.

(i) Si on considère comme « normale » toute valeur se situant à deux écarts-types ou moins de la moyenne, quelles sont les limites « normales » ? Avec ce critère, la température maximale observée de 100,8 est-elle anormale ?

(ii) Déterminer les cotes Z de tous les sujets et faites-en un diagramme de points. Il y a trois personnes qui sont visiblement éloignées du groupe. Si on conclut que ces trois personnes sont malades (ou ont mal lu le thermomètre), on doit les éliminer avant d'établir des limites de « normalité ». Quelles sont les limites une fois ces données éliminées ?

e) Y a-t-il une différence de température entre hommes et femmes ? Répondez par des mesures descriptives et par des graphiques si ceux-ci sont révélateurs. Éliminez d'abord la plus grande des données.

f) Y a-t-il une relation entre la température (Y) et le nombre de battements du cœur (X) ?

(i) Faites un nuage de points et déterminez la droite de régression et le coefficient de corrélation.

(ii) Est-ce que la relation entre X et Y semble différente pour les hommes et les femmes ?

1.21 [*Données du tableau* A.3] Le tableau A.3 en annexe présente des données sur 29 sujets desquels on a obtenu une mesure de la grosseur du cerveau ainsi que certaines mesures d'aptitude mentale.

a) Faites un graphique permettant de voir si la variable P, le score de performance, dépend de la taille du cerveau. Vous devriez constater que la relation, si elle existe, est plutôt faible.

b) Utilisez un symbole différent pour les femmes et pour les hommes. Que constatez-vous ?

c) Vérifiez que la taille du cerveau est corrélée avec la taille de la personne.

d) Étant donné la constatation faite en c), déterminez la relation entre P et *IRM*/*taille* (cette dernière variable ajuste le poids du cerveau en l'exprimant comme proportion de la taille de la personne). Y a-t-il une amélioration dans le coefficient de corrélation ?

1.22 [*Données du tableau* A.6] Le tableau A.6 présente des données démographiques et économiques sur 96 pays.

a) Montrez à l'aide d'un diagramme de points la distribution du PNB pour les pays du moyen et proche orient. Quels sont ces trois pays dont le PNB est nettement supérieur à celui du groupe ? Expliquez ce que ces pays ont de particulier. Calculez la moyenne et l'écart-type des *autres* pays du groupe et servez-vous en pour déterminer si les trois pays en question ressortent vraiment du

groupe. Confirmer le caractère distinct de ces pays en les comparant aux autres par rapport aux autres variables. Finalement, si vous trouvez que ces trois pays sont suffisamment distincts des autres, placez-les dans une classe à part.

b) Dans les pays de l'extrême orient, identifiez les trois qui devraient être retirés du groupe.

c) Faites un graphique pour montrer la relation entre le PNB et l'espérance de vie des hommes (MVie). Vous verrez qu'elle est loin d'être linéaire. Essayez d'expliquer pourquoi le nuage a cette forme.

d) Déterminez maintenant la droite de régression permettant de prédire l'espérance de vie des hommes à partir du logarithme du PNB (c'est-à-dire, construisez une colonne contenant x = ln *PNB*, puis faites une régression de l'espérance de vie (Y) sur x. Estimez l'espérance de vie des hommes d'un pays dont le PNB est de 1000 $ par habitant.

e) La relation entre le taux de mortalité infantile et le PNB ne semble pas non plus linéaire. Supposons que la relation entre y, le taux de mortalité et x, le PNB, est de la forme suivante : $y = \alpha x^{\beta}$. Il s'ensuit que $\ln y = \ln \alpha + \beta \ln x$. Examinez alors le lien entre le logarithme du PNB et le logarithme du taux de mortalité. Vérifiez visuellement que la relation est linéaire et calculez le coefficient de corrélation. Estimez le taux de mortalité d'un pays dont le PNB est de 1000 $ par habitant.

f) Présentez graphiquement la distribution des variables Fvie et Mvie). Identifiez les données extrêmes.

g) Évaluez la relation entre FVie (Y) et MVie (x). Interprétez le fait que le coefficient de MVie est supérieur à 1.

1.23 [*Données du tableau* A.7] Le tableau A.7 présente des données économiques sur 46 grandes villes du monde.

a) Faites un graphique montrant la relation entre le coût des produits (x) et les salaires (y).

b) Considérer les différences $y_i - \overline{y}$ ainsi que les différences $y_i - \hat{y}_i$, où $\hat{y}_i$ est la prédiction de y à partir de x_i : $\hat{y}_i = b_0 + b_1 x_i$. Calculez l'écart-type de chacune de ces deux séries. Considérer le sens de ces écarts pour expliquer pourquoi la deuxième série ne peut pas être plus dispersée que la première.

1.24 [*Données du tableau* A.8] Présentez des données ou des tableaux qui confirment ou infirment les propositions suivantes :

a) Ceux qui croient à l'évolution n'ont pas tendance à croire à l'astrologie

b) Ceux qui vont régulièrement à l'église préfèrent généralement épouser des coreligionnaires.

c) Ceux qui croient à l'astrologie ont tendance à croire à la valeur de prédiction de la ligne de vie.